Interval Estimators for Inequality Measures Using Grouped Data

Dilanka S. Dedduwakumara(✉) and Luke A. Prendergast

Department of Mathematics and Statistics, La Trobe University, Melbourne, Australia
18748354@students.latrobe.edu.au

Abstract. Income inequality measures are often used as an indication of economic health. How to obtain reliable confidence intervals for these measures based on sampled data has been studied extensively in recent years. To preserve confidentiality, income data is often made available in summary form only (i.e. histograms, frequencies between quintiles, etc.). In this paper, we show that good coverage can be achieved for bootstrap and Wald-type intervals for quantile-based measures when only grouped (binned) data are available. These coverages are typically superior to those that we have been able to achieve for intervals for popular measures such as the Gini index in this grouped data setting. To facilitate the bootstrapping, we use the Generalized Lambda Distribution and also a linear interpolation approximation method to approximate the underlying density. The latter is possible when groups means are available. We also apply our methods to real data sets.

Keywords: Histograms · Inequality measures · Bootstrap confidence intervals · Generalized Lambda Distribution

1 Introduction

Income data are generally made available in binned formats by governing bodies to preserve the confidentiality of the individual participants. Obtaining inferences from such summary information has been recently discussed by Deduwakumara and Prendergast (2018), in the context of obtaining confidence intervals for quantiles using estimates of the underlying distribution using grouped data. As we will show in what follows, we can obtain reliable confidence intervals for some inequality measures using bootstrap and Wald-type approaches.

Motivated by these findings, we compare the interval estimators for inequality measures when the data are available in grouped form only. For comparison, we use the well-known Gini, Theil and Atkinson indices and the newly proposed quantile ratio index (Prendergast and Staudte 2018). We begin by introducing these measures before discussing some distribution estimation strategies in Sect. 3. In Sect. 4, we report findings of simulations for interval estimators of the inequality measures. Two real data examples are presented in Sect. 5, followed by a brief discussion in Sect. 6.

H. Nguyen (Ed.): RSSDS 2019, CCIS 1150, pp. 238–252, 2019.
https://doi.org/10.1007/978-981-15-1960-4_17

2 Some Inequality Measures

Let f, F and Q denote the density, distribution and quantile functions respectively for the population of interest. For $p \in [0,1]$, let $x_p = Q(p) = F^{-1}(p)$ denote the p-th quantile. We find it convenient to consider continuous probability distributions to model incomes while acknowledging that, in practice, a population of incomes has a finite number, N, of individuals. Let $x_1, \ldots, x_n$ denote a simple random sample of incomes from the population and let $\widehat{x}_p$ be the estimated p-th quantile.

2.1 Gini Index

Suppose $X \sim F$ where X represents a randomly chosen income from the population and let $\mu = E(X)$ denote mean income. Easily the most commonly used inequality measure is the Gini index (Gini 1914), which measures the deviation of the income distribution from perfect equality. It can be defined as,

$$G = 1 - \frac{1}{\mu} \int_0^\infty [1 - F(x)]^2 \, dx$$

with $G \in [0,1]$. Here, $G = 1$ indicates that one individual holds all wealth (e.g. one individual with income greater than zero) and $G = 0$ represents the equality of incomes for all. The Gini index can be estimated for a simple random sample of size n, with the ordered values of $x_1, \ldots, x_n$ by,

$$\hat{G} = \frac{2 \sum_i i x_i}{n \sum_i x_i} - \frac{n+1}{n}.$$

For more details on the Gini index and estimation see, for example, Dixon *et al.* (1988) and Damgaard and Weiner (2000).

2.2 Theil Index

Based on information theory, Theil (1967) proposed an entropy-based measure which is defined to be

$$T = \int_0^\infty \left(\frac{x}{\mu}\right) \log\left(\frac{x}{\mu}\right) f(x) \, dx$$

where $T \in [0, \infty)$. In practice where a population consists of finite number of N incomes, the upper bound is $\ln(N)$. The Theil index can be estimated by

$$\widehat{T} = \frac{1}{n} \sum_i \frac{x_i}{\bar{x}} \ln\left(\frac{x_i}{\bar{x}}\right)$$

where $\bar{x}$ is the sample mean and where $\widehat{T} \in [0, \ln(n)]$. Further properties of the Theil index can be found in Theil (1967), Allison (1978) and Shorrocks (1980).

2.3 Atkinson Index

The Atkinson index was initially introduced by Atkinson (1970). This measure depends on the sensitivity parameter, ϵ $(0 < \epsilon < \infty)$, which represents the level of inequality aversion. As this parameter increases, more weight is shifted to the distribution at the lower end and vice versa. It is defined as

$$A = 1 - \left[\int_0^\infty \left(\frac{x}{\mu} \right)^{1-\epsilon} f(x)\; dx \right]^{\frac{1}{1-\epsilon}}$$

where $A \in [0, 1]$.

Atkinson values represent the proportion of total income that would be needed to achieve an equal level of social welfare if incomes were perfectly distributed. Depending on the value of ϵ, the sample estimate is

$$\hat{A} = \begin{cases} 1 - \dfrac{1}{\bar{x}} \left(\dfrac{1}{n} \sum_i x_i^{1-\epsilon} \right)^{\frac{1}{1-\epsilon}}, & \text{for} \quad 0 \le \epsilon < 1 \\ 1 - \dfrac{1}{\bar{x}} \left(\prod_i x_i \right)^{\frac{1}{n}}, & \text{for} \quad \epsilon = 1 \end{cases}$$

We use the value of $\epsilon = 0.5$ for our analysis which is the default value used in the package `ineq` (Zeileis 2014) in `R` software (R Core Team 2017). More details for the Atkinson index can be found in Atkinson (1970), Biewe and Jenkins (2006) and Shorrocks (1980).

2.4 Quantile Ratio Index

Prendergas and Staude (2018, 2019) introduced the quantile ratio index (QRI) which uses the ratio of symmetric quantiles and which is simpler than similarly defined inequality measures given by Prendergast and Staudte (2016b). The QRI is denoted as

$$I = 1 - \int_0^1 \frac{x_{p/2}}{x_{1-p/2}}\; dp = 1 - \int_0^1 R(p)\; dp$$

where $I \in [0, 1]$. Note that $R(p)$ is the ratio of symmetric quantiles so that I can be seen to be based on the average ratio of incomes chosen symmetrically from the poorer and richer halves of the incomes respectively. For a suitably large J, I is estimated as $J^{-1} \sum_j \left[1 - \widehat{R}(p_j) \right]$ where $p_j = (j - 1/2)/J$ and $\widehat{R}(p_j)$ is the ratio of the estimated $(p_j/2)$-th and $(1 - p_j/2)$-th quantiles. Prendergast and Staudte (2018) show that $J = 100$ is large enough to obtain good estimates of I and so this will be our choice in what follows.

3 Density Estimation Methods

We now consider two methods for estimating the density from grouped data. The first requires bins and frequencies, and the second also requires the bin means. The methods were used by Dedduwakumara and Prendergast (2018) to obtain intervals for quantiles from histograms.

3.1 GLD Estimation Method

Due to flexibility in approximating a wide range of distributions, the Generalized Lambda Distribution (GLD) is commonly used and particularly favoured in fields such as economics and finance. Defined in terms of its quantile function, several parameterizations for the GLD exist. Following is the FKML parameterization for the GLD given by Freimer *et al.* (1988) which is often favoured since it is defined for all parameter choices, with the only restriction being that the scale parameter must be greater than zero. The GLD quantile function is

$$Q(p) = \lambda + \frac{1}{\eta}\left[\frac{(p^\alpha - 1)}{\alpha} - \frac{(1-p)^\beta - 1}{\beta}\right]. \tag{1}$$

The GLD has been used in different contexts to obtain various interval estimators (e.g. Su 2009; Prendergast and Staudte 2016a) when the full data set is available. However, using the percentile matching methods presented by Karian and Dudewicz (1999) and Tarsitano (2005), the GLD parameters can still be estimated when data is in grouped format with frequencies and bins. This method is available in the `bda` package (Wang 2015).

3.2 Linear Interpolation Method

The linear interpolation method was proposed by Lyon *et al.* (2016) as a method of estimating the underlying distribution of binned data when the group (bin) means are also available. Within each bin, a linear density is estimated using the lower and upper bounds of the bin and the associated mean, and the final bin is fitted with an unbounded exponential tail. The slope of the linear density is determined by the mean in relation to the bin midpoint. Closed form solutions for the density and the quantile functions are extensively provided by Lyon *et al.* (2016) and following is a summary of the density results.

Assume there are J intervals in the grouped data bounded by $[a_{j-1}, a_j), j = 1, \ldots, J$ where $a_0 > -\infty$ and $a_J = \infty$. Let the midpoint, mean and relative frequency of the jth bin be denoted by x_j^c, $\bar{x}_j$ and $\widehat{f}_j$. The linear density for the jth bin is

$$h_j(x) = \alpha_j + \beta_j x, \qquad x \in [a_{j-1}, a_j) \tag{2}$$

where the estimates of α_j, β_j are given by,

$$\widehat{\beta}_j = \widehat{f}_j \frac{12(\bar{x}_j - x_j^c)}{(a_j - a_{j-1})^3}, \quad \widehat{\alpha}_j = \frac{\widehat{f}_j}{a_j - a_{j-1}} - \widehat{\beta}_j x_j^c. \tag{3}$$

The density estimate for the final unbounded interval using an exponential tail is provided by,

$$h_J(x) = \frac{\eta}{\lambda} \exp\left\{ -\frac{(x - a_{J-1})}{\lambda} \right\} \tag{4}$$

where $\widehat{\eta} = \widehat{f_J}$ and $\widehat{\lambda} = \bar{x}_J - a_{J-1}$.

4 Interval Estimators Using Grouped Data

In this section, we propose and describe our bootstrap and Wald-type methods to produce intervals for inequality measures using grouped information. The variance of the QRI estimator depends on the underlying income distribution density function applied to income quantiles (Prendergast and Staudte 2018). Therefore, provided we can obtain good estimates of the density from grouped data, then the QRI is well-suited to obtaining Wald-type intervals in this setting. Aside from bootstrapping, to obtain the variance of, for example, the Gini index, it is common to use the jackknife approach or other methods that require the full data set. Consequently, obtaining an approximation to the variances for the Gini, Thiel and Atkinson measure estimators from grouped data is not straightforward and therefore an area for further research.

For the bootstrapping procedure, we obtain the bootstrap samples from the estimated quantile function arising from the estimated GLD or linear interpolation densities. We then use the percentile bootstrap interval described below. While there are other bootstrap methods available that often have improved performance over the percentile method, they require the full data set and it is not immediately clear on how to use them when data is only available in grouped format; e.g. the bootstrap t interval requires the variance of the estimator, the BCa method (Efron 1987) and Efron's ABC method (Diciccio and Efron 1992) requires the full sample data to calculate the acceleration parameter. However, we did try a variation of the bootstrap t interval whereby the α parameter was estimated as usual, but where the estimate and its standard error were also approximated from the bootstrap samples given the lack of the full data set. Coverages were usually no better, and often worse than those for the percentile approach so we do not present them in what follows for brevity. Further variations of bootstrap methods to accommodate the lack of the full data set may result in improved results and this is an area for future research.

Bootstrap Confidence Intervals. In the following algorithm, we describe the estimation of percentile bootstrap confidence intervals in detail.

Step 1: Estimate the GLD and linear interpolation densities using available summary information of bin points and frequencies (and bin means for the linear interpolation approach).

Step 2: Take 500 bootstrap samples of size n using the estimated quantile functions from the two estimation methods using the inverse transform sampling method. That is, randomly generate n numbers, $y_1, \ldots, y_n$ in $[0, 1]$ from the uniform distribution and then the ith observation for the jth bootstrap is $y_{ji} = \widehat{Q}(y_i)$ where $\widehat{Q}$ is the estimated quantile function.

Step 3: Construct the percentile bootstrap 95% confidence intervals by taking the 2.5% and 97.5% quantiles of the 500 bootstrapped estimates of the inequality measures.

For the GLD method, we consider the available bin points as the empirical percentiles in the percentile matching method, providing the estimated parameters for the GLD. By using the GLD quantile function (Sect. 3.1) and the estimated parameters, we can easily take the bootstrap samples using the inverse transform sampling method as in Step 2. For the linear interpolation approach, we use the following two quantile functions to generate data depending on the value of p (Lyon *et al.* 2016). For the bounded interval of $[a_{j-1}, a_j)$, the following quantile function is used for $p \in [0, 1)$ is,

$$\widehat{x}_p = \frac{-\widehat{\alpha}_j + \sqrt{2\widehat{\beta}_j p + \widehat{C}_j}}{\widehat{\beta}_j} \tag{5}$$

where, $\widehat{C_j} = [\widehat{\alpha}_j^2 - 2\widehat{\beta}_j \widehat{F}_{j-1} + 2\widehat{\beta}_j \widehat{\alpha}_j a_{j-1} + \widehat{\beta}_j^2 (a_{j-1})^2]$, $\widehat{\beta_j}$ and $\widehat{\alpha_j}$ as in (3).

Further the fitted exponential tail yields the following quantile function when the cumulative relative frequency up to final (Jth) interval is denoted by $\widehat{F}_J$,

$$\widehat{x}_p = a_{J-1} - \widehat{\lambda} \ln \left(1 - \frac{p - \widehat{F}_{J-1}}{\widehat{\eta}}\right). \tag{6}$$

Wald-Type Confidence Intervals for the QRI. Obtaining confidence intervals for the QRI from full data sets is studied by Prendergast and Staudte (2018). The variance of the estimator depends on the density function and quantiles. Therefore, given a good estimation of the density which in turn would be expected to give good estimates to quantiles, QRI intervals from grouped data are possible.

The $(1-\alpha) \times 100$ confidence interval for I is given by $\hat{I} \pm z_{1-\alpha/2}\sqrt{\text{Var}(\hat{I})}$, where $\text{Var}(\hat{I})$ is adopted from Prendergast and Staudte (2018) where we use $J = 100$. Here, $z_{1-\alpha/2}$ is the $1-\alpha/2$ percentile from the standard normal distribution. $\text{Var}(\hat{I})$ consists of the variances and co-variances terms of ratios of symmetrically chosen quantiles (see Prendergast and Staudte 2018). We then require estimates for population quantiles and density function. As described earlier, first we estimate the underlying density and quantile functions using the GLD and linear interpolation methods. Then those estimated quantile functions can be used to estimate the symmetrically chosen quantiles.

5 Simulations and Examples

We begin by reporting our findings for simulation studies conducted with a variety of distributions before considering real data examples.

5.1 Simulations

To assess coverage, we consider the lognormal distribution with $\mu = 0$ and $\sigma = 1$ and the Singh-Maddala distribution with parameter values $a = 1.6971$, $b = 87.6981$ and $q = 8.3679$ where these parameters were from fitted US family incomes reported by McDonald (1984). We also consider the Dagum distribution with the parameter choices of $a = 4.273$ $b = 14.28$ and $p = 0.36$ which were used in Kleiber (2008) and were estimated from fitted US family incomes in 1969. The χ^2_2, Pareto type II distribution with scale one and shape equal to two and the exponential distribution with rate one were also considered. Table 1 provides the population inequality values of each measure.

Table 1. True values of inequality measures for each distribution.

F	Gini	Theil	Atkinson	I
Lognormal	0.520	0.500	0.221	0.664
Singh-Maddala	0.355	0.206	0.106	0.579
Dagum	0.335	0.191	0.097	0.548
χ^2_2	0.500	0.423	0.215	0.702
Pareto (2)	0.667	1.000	0.383	0.740
Exponential (1)	0.500	0.423	0.215	0.702
Weibull (10)	0.067	0.007	0.004	0.167

From Table 2 for quintile-grouped data and using the linear interpolation method, intervals for I produces coverage probabilities close to the nominal level of 0.95 together with narrow mean width for all settings and with both bootstrap and the Wald-type intervals. Given that the computation of the interval is much more efficient for the Wald-type interval, there does not appear to be an advantage for using the bootstrap. However, for the Gini, Theil and Atkinson

Table 2. Empirical coverage probabilities and average widths (in brackets) of Bootstrapped interval estimates of inequality measures from quintiles estimated using linear interpolation method at nominal level 95%, each based on 1000 replications and 500 bootstrap repetitions.

F	n	Bootstrap				Wald-type
		Gini	Theil	Atkinson	I	I
Lognormal	50	0.788 (0.164)	0.734 (0.327)	0.785 (0.129)	0.947 (0.162)	0.968 (0.163)
	100	0.813 (0.119)	0.761 (0.250)	0.804 (0.097)	0.960 (0.112)	0.965 (0.112)
	250	0.837 (0.075)	0.720 (0.161)	0.813 (0.062)	0.967 (0.069)	0.962 (0.070)
	500	0.840 (0.054)	0.650 (0.115)	0.798 (0.045)	0.955 (0.048)	0.956 (0.049)
Singh-Maddala	50	0.909 (0.128)	0.921 (0.151)	0.911 (0.072)	0.948 (0.165)	0.949 (0.164)
	100	0.925 (0.091)	0.927 (0.108)	0.914 (0.052)	0.933 (0.114)	0.959 (0.116)
	250	0.940 (0.058)	0.948 (0.069)	0.938 (0.034)	0.933 (0.072)	0.947 (0.072)
	500	0.946 (0.041)	0.952 (0.049)	0.946 (0.024)	0.941 (0.050)	0.948 (0.051)
Dagum	50	0.902 (0.128)	0.886 (0.143)	0.869 (0.069)	0.939 (0.169)	0.946 (0.168)
	100	0.914 (0.093)	0.902 (0.105)	0.904 (0.051)	0.952 (0.117)	0.951 (0.118)
	250	0.902 (0.059)	0.878 (0.067)	0.893 (0.033)	0.940 (0.073)	0.948 (0.074)
	500	0.925 (0.042)	0.891 (0.048)	0.918 (0.024)	0.943 (0.052)	0.954 (0.052)
χ^2_2	50	0.930 (0.158)	0.939 (0.285)	0.931 (0.126)	0.954 (0.170)	0.964 (0.170)
	100	0.930 (0.111)	0.933 (0.204)	0.930 (0.090)	0.955 (0.117)	0.952 (0.118)
	250	0.938 (0.071)	0.939 (0.131)	0.939 (0.058)	0.951 (0.072)	0.952 (0.073)
	500	0.948 (0.050)	0.950 (0.093)	0.946 (0.041)	0.945 (0.051)	0.960 (0.051)
Pareto (2)	50	0.633 (0.172)	0.391 (0.490)	0.603 (0.177)	0.968 (0.163)	0.969 (0.162)
	100	0.637 (0.121)	0.351 (0.373)	0.590 (0.131)	0.970 (0.112)	0.971 (0.113)
	250	0.571 (0.077)	0.172 (0.242)	0.484 (0.084)	0.949 (0.069)	0.959 (0.070)
	500	0.500 (0.054)	0.083 (0.173)	0.362 (0.060)	0.973 (0.048)	0.961 (0.049)
Exponential (1)	50	0.916 (0.158)	0.934 (0.288)	0.921 (0.126)	0.939 (0.169)	0.965 (0.170)
	100	0.929 (0.111)	0.938 (0.204)	0.924 (0.090)	0.952 (0.116)	0.966 (0.118)
	250	0.936 (0.071)	0.949 (0.131)	0.935 (0.058)	0.929 (0.072)	0.962 (0.073)
	500	0.943 (0.050)	0.945 (0.093)	0.947 (0.041)	0.961 (0.050)	0.963 (0.051)

measures, the coverages are comparatively weaker but improves as the sample size increases for most of the distributions.

Table 3 shows that the intervals based on the GLD and quintiles for the Gini, Theil and Atkinson measures have poor coverage. Coverages are typically very good for the QRI intervals, albeit more conservative than those using the linear interpolation method. However, coverages become low for the lognormal suggesting that quintiles do not provide enough information to get a good approximation using the GLD.

Table 3. Empirical coverage probabilities and average widths (in brackets) of Bootstrapped interval estimates of inequality measures from quintiles estimated using GLD method at nominal level 95% for, each based on 1000 replications and 500 bootstrap repetitions.

F	n	Bootstrap				Wald-type
		Gini	Theil	Atkinson	I	I
Lognormal	50	0.495 (0.168)	0.406 (0.387)	0.598 (0.150)	0.967 (0.173)	0.974 (0.172)
	100	0.446 (0.117)	0.366 (0.260)	0.510 (0.101)	0.975 (0.126)	0.971 (0.105)
	250	0.373 (0.071)	0.269 (0.141)	0.453 (0.059)	0.899 (0.085)	0.924 (0.065)
	500	0.271 (0.049)	0.165 (0.090)	0.359 (0.039)	0.661 (0.063)	0.713 (0.046)
Singh-Maddala	50	0.862 (0.134)	0.937 (0.151)	0.953 (0.090)	0.979 (0.168)	0.989 (0.180)
	100	0.783 (0.094)	0.920 (0.107)	0.930 (0.063)	0.984 (0.119)	0.973 (0.125)
	250	0.735 (0.060)	0.918 (0.068)	0.911 (0.040)	0.974 (0.075)	0.955 (0.078)
	500	0.646 (0.042)	0.887 (0.048)	0.803 (0.028)	0.965 (0.054)	0.925 (0.056)
Dagum	50	0.844 (0.133)	0.955 (0.140)	0.988 (0.085)	0.990 (0.174)	0.988 (0.192)
	100	0.759 (0.094)	0.909 (0.099)	0.991 (0.060)	0.991 (0.123)	0.981 (0.132)
	250	0.561 (0.060)	0.799 (0.063)	0.982 (0.038)	0.982 (0.079))	0.959 (0.083)
	500	0.299 (0.042)	0.575 (0.045)	0.981 (0.027)	0.967 (0.057)	0.941 (0.059)
χ^2_2	50	0.652 (0.169)	0.544 (0.359)	0.749 (0.158)	0.980 (0.170)	0.989 (0.172)
	100	0.583 (0.121)	0.488 (0.269)	0.663 (0.111)	0.971 (0.117)	0.978 (0.118)
	250	0.605 (0.073)	0.512 (0.147)	0.666 (0.065)	0.970 (0.073)	0.979 (0.073)
	500	0.568 (0.051)	0.467 (0.096)	0.624 (0.044)	0.974 (0.051)	0.969 (0.051)
Pareto (2)	50	0.558 (0.237)	0.508 (1.029)	0.609 (0.289)	0.973 (0.161)	0.989 (0.161)
	100	0.579 (0.197)	0.549 (1.056)	0.607 (0.251)	0.971 (0.111)	0.977 (0.111)
	250	0.626 (0.152)	0.647 (0.982)	0.663 (0.201)	0.968 (0.069)	0.972 (0.069)
	500	0.650 (0.123)	0.697 (0.903)	0.687 (0.169)	0.976 (0.048)	0.977 (0.049)
Exponential (1)	50	0.653 (0.172)	0.559 (0.388)	0.722 (0.163)	0.973 (0.169)	0.980 (0.171)
	100	0.589 (0.119)	0.513 (0.259)	0.667 (0.110)	0.970 (0.117)	0.983 (0.118)
	250	0.578 (0.074)	0.483 (0.151)	0.651 (0.066)	0.982 (0.073)	0.973 (0.073)
	500	0.561 (0.051)	0.470 (0.095)	0.615 (0.044)	0.973 (0.051)	0.969 (0.051)

When the data is summarised in deciles rather than quintiles (i.e. more bins and more information), Table 4 shows improved coverage is achieved with the GLD method. However, coverage is still poor for the Gini, Theil and Atkinson measures when compared to the good coverages achieved for the QRI. Again, the similar coverages for the bootstrap and Wald-type intervals suggest that the Wald-type is a good choice since it is simple and quick to compute.

Table 4. Empirical coverage probabilities and average widths (in brackets) of Bootstrapped interval estimates of inequality measures from deciles estimated using GLD method at nominal level 95% for, each based on 1000 replications and 500 bootstrap repetitions.

F	n	Bootstrap				Wald-type
		Gini	Theil	Atkinson	I	I
Lognormal	50	0.754 (0.262)	0.733 (0.963)	0.762 (0.273)	0.926 (0.156)	0.948 (0.156)
	100	0.789 (0.209)	0.787 (0.892)	0.781 (0.227)	0.943 (0.108)	0.953 (0.109)
	250	0.760 (0.152)	0.761 (0.749)	0.756 (0.173)	0.938 (0.068)	0.943 (0.068)
	500	0.740 (0.113)	0.744 (0.585)	0.730 (0.130)	0.927 (0.048)	0.920 (0.048)
Singh-Maddala	50	0.791 (0.148)	0.760 (0.248)	0.769 (0.103)	0.912 (0.160)	0.958 (0.161)
	100	0.781 (0.102)	0.756 (0.167)	0.747 (0.068)	0.922 (0.111)	0.965 (0.113)
	250	0.786 (0.060)	0.748 (0.083)	0.715 (0.037)	0.941 (0.070)	0.954 (0.071)
	500	0.756 (0.041)	0.706 (0.052)	0.660 (0.025)	0.945 (0.050)	0.955 (0.050)
Dagum	50	0.735 (0.146)	0.631 (0.222)	0.740 (0.101)	0.898 (0.163)	0.937 (0.167)
	100	0.744 (0.099)	0.632 (0.138)	0.733 (0.067)	0.941 (0.115)	0.956 (0.118)
	250	0.709 (0.060)	0.564 (0.074)	0.685 (0.039)	0.957 (0.073)	0.960 (0.074)
	500	0.710 (0.042)	0.499 (0.047)	0.681 (0.027)	0.957 (0.052)	0.949 (0.052)
χ_2^2	50	0.807 (0.202)	0.783 (0.551)	0.845 (0.196)	0.941 (0.165)	0.958 (0.166)
	100	0.775 (0.141)	0.736 (0.392)	0.803 (0.134)	0.954 (0.115)	0.952 (0.116)
	250	0.799 (0.084)	0.763 (0.216)	0.779 (0.077)	0.969 (0.071)	0.959 (0.072)
	500	0.753 (0.057)	0.714 (0.136)	0.742 (0.050)	0.970 (0.050)	0.957 (0.051)
Pareto (2)	50	0.747 (0.283)	0.682 (1.374)	0.775 (0.355)	0.930 (0.159)	0.948 (0.160)
	100	0.787 (0.236)	0.745 (1.414)	0.800 (0.312)	0.945 (0.110)	0.939 (0.111)
	250	0.815 (0.185)	0.817 (1.370)	0.839 (0.258)	0.935 (0.068)	0.911 (0.069)
	500	0.812 (0.149)	0.856 (1.244)	0.845 (0.214)	0.905 (0.048)	0.928 (0.048)
Exponential (1)	50	0.802 (0.200)	0.762 (0.537)	0.822 (0.192)	0.920 (0.165)	0.953 (0.167)
	100	0.830 (0.142)	0.780 (0.395)	0.826 (0.135)	0.943 (0.115)	0.959 (0.116)
	250	0.785 (0.087)	0.743 (0.232)	0.781 (0.080)	0.968 (0.071)	0.957 (0.072)
	500	0.756 (0.057)	0.720 (0.139)	0.748 (0.051)	0.972 (0.050)	0.953 (0.051)

In Fig. 1 we look at what happens to estimates using the linear interpolation method for each measure (e.g. an estimate based on a bootstrap sample) as skew increases. In this case, we use the lognormal distribution while increasing the σ parameter from 0.5 to 2. The estimates are centered according to the true value so a value of zero indicates a perfect estimate. We exclude the Theil index from the analysis since its upper bound is unrestricted. As the distribution becomes more skewed, the Gini and Atkinson estimators have an increase in bias and variability whereas the quantile-based measure (I) indicates smaller variability

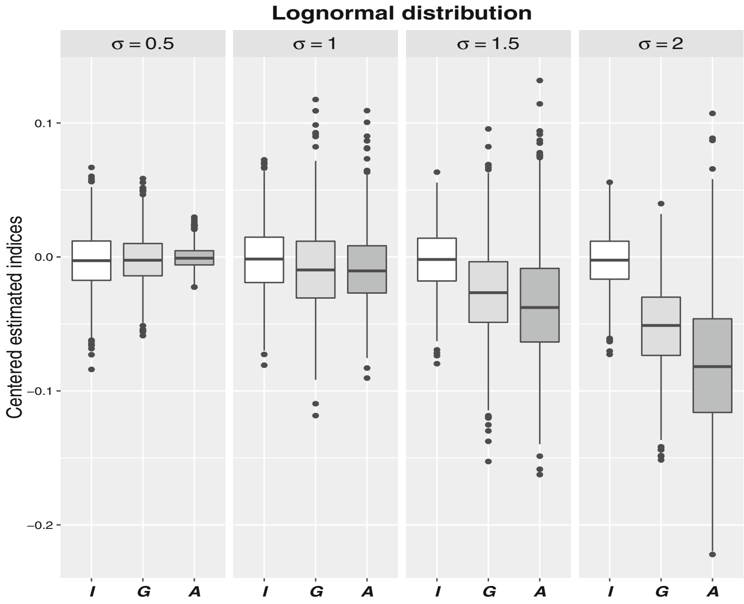

Fig. 1. Boxplots of 1000 centered (with respect to the true values) simulated estimates of inequality measures from quintiles, estimated using linear interpolation method from the Lognormal distribution with mean 0 and various standard deviation values where $n = 250$

and smaller bias throughout for all of the choices of σ. This helps to explain why the coverages are poor for the Gini and Atkinson measures.

6 Applications

6.1 Example 1: Household Income Reported with Group Means

In this example, we present household income data reported with group means by the Survey of Consumer Finances and Expenditures carried out by the Macquarie University and the University of Queensland which can be found in Podder (1972) and Kakwani and Podder (1976). The data is summarised in Table 5.

The confidence intervals produced by 500 bootstrapped samples using the linear interpolation (LI) and GLD methods are given in Table 6. As the final interval is unbounded, we arbitrarily set the upper limit of that bin to $500,000. As can be seen, the confidence intervals and the estimates generated by the two methods are similar.

Table 5. Australian household income data for 1967-68

Income	Number of households	Mean income
Below $1000	310	674.39
$1000–$2000	552	1426.10
$2000–$3000	1007	2545.79
$3000–$4000	1193	3469.35
$4000–$5000	884	4470.33
$5000–$6000	608	5446.60
$6000–$7000	314	6460.93
$7000–$8000	222	7459.14
$8000–$9000	128	8456.66
$9000–$11000	112	9788.38
$11000 and over	110	15617.69

Table 6. Interval and point estimates of the inequality measures generated using the linear interpolation (LI) and GLD methods for the data presented in Table 7.

Method	Bootstrap				Wald-type
	Gini	Theil	Atkinson	I	I
LI	0.319	0.178	0.088	0.509	0.510
	(0.311, 0.327)	(0.168, 0.188)	(0.084, 0.092)	(0.503, 0.517)	(0.502, 0.517)
GLD	0.329	0.177	0.104	0.519	0.521
	(0.321, 0.337)	(0.165, 0.190)	(0.098, 0.109)	(0.512, 0.528)	(0.513, 0.529)

6.2 Example 2: Comparison of Equalized Disposable Household Income Data

In this example, we compare two assumed-independent income distributions reported in deciles from ABS (2011) (see Table 7) to assess whether the income inequality measures of the two distributions are significantly different from one another. It is simple to adapt the previous intervals to the two-sample setting. For example, for the bootstrap approach we simply estimate the difference at each iteration and then form the interval by taking percentiles from the bootstrapped differences. For the Wald-type approach we can get the variance of the difference as a sum of the variances for each estimator of the QRI. For estimation purposes, the highest income has been considered as $5000 for both years.

From Table 8, it can be seen that all intervals for the difference in the measures do not include zero. These intervals then suggest that income inequality has change over the years. We can conclude that inequality of the equalized disposable household income in Western Australia has been significantly increased from 1996-97 to 2009-10.

Table 7. Equalized disposable household income at top of selected percentiles ($) in Western Australia.

Percentile	1996-97	2009-10
10th	263	347
20th	311	454
30th	364	565
40th	434	663
50th	518	770
60th	586	882
70th	665	1071
80th	778	1296
90th	955	1652

Table 8. Point and interval estimates of inequality measures generated using GLD method for Equalized disposable household income in Western Australia presented in Table 7

Year		Bootstrap				Wald-type
		Gini	Theil	Atkinson	I	I
1996-97	Average Est.	0.262	0.107	0.053	0.488	0.489
	CI	(0.253, 0.271)	(0.099, 0.115)	(0.049, 0.057)	(0.473, 0.503)	(0.483, 0.496)
2009-10	Average Est.	0.326	0.174	0.083	0.538	0.538
	CI	(0.318, 0.334)	(0.163, 0.185)	(0.079, 0.088)	(0.528, 0.548)	(0.531, 0.545)
Difference	Average Est.	0.064	0.067	0.030	0.050	0.049
	CI	(0.051, 0.077)	(0.054, 0.08)	(0.025, 0.037)	(0.032, 0.07)	(0.040, 0.058)

7 Discussion

To preserve confidentiality, it is common for income data to be summarised in grouped format. We therefore considered interval estimators for several measures, including the popular Gini index and a newly proposed quantile-based measure, the QRI. Since grouped data contains bin boundaries and frequencies (and therefore quantile estimates of the data), the QRI is naturally suited to this setting. We showed that bootstrap intervals and a Wald-type interval, both using estimated densities form the grouped data, had typically excellent coverage (i.e. close to nominal). The other measures, however, often had intervals with poor coverage. Further research could include consideration of how to get good approximations to the variances of the Gini, Theil and Atkinson estimators when dealing with grouped data. This was possible for the QRI since the variance of the estimator can be approximated using the estimated density function. For the other measures it is not so straightforward. In summary, when faced with grouped data, if confidence intervals are needed then the QRI is a good option for measuring inequality.

References

ABS: Household income and income distribution, Australian bureau of statistics report 6523.0, 2009–10 (2011)

Allison, P.D.: Measures of inequality. Am. Sociol. Rev. 865–880 (1978)

Atkinson, A.B.: On the measurement of inequality. J. Econ. Theory **2**(3), 244–263 (1970)

Biewen, M., Jenkins, S.P.: Variance estimation for generalized entropy and atkinson inequality indices: the complex survey data case. Oxf. Bull. Econ. Stat. **68**(3), 371–383 (2006)

Damgaard, C., Weiner, J.: Describing inequality in plant size or fecundity. Ecology **81**(4), 1139–1142 (2000)

Dedduwakumara, D.S., Prendergast, L.A.: Confidence intervals for quantiles from histograms and other grouped data. Commun. Stat. Simul. Comput. 1–14 (2018)

Diciccio, T., Efron, B.: More accurate confidence intervals in exponential families. Biometrika **79**(2), 231–245 (1992)

Dixon, P.M., Weiner, J., Mitchell-Olds, T., Woodley, R.: Erratum to 'bootstrapping the gini coefficient of inequality'. Ecology **69**(4), 1307 (1988)

Efron, B.: Better bootstrap confidence intervals. J. Am. Stat. Assoc. **82**(397), 171–185 (1987)

Freimer, M., Kollia, G., Mudholkar, G.S., Lin, C.T.: A study of the generalized tukey lambda family. Commun. Stat. Theory Methods **17**(10), 3547–3567 (1988)

Gini, C.: Sulla misura della concentrazione e della variabilità dei caratteri (1914)

Kakwani, N.C., Podder, N.: Efficient estimation of the Lorenz curve and associated inequality measures from grouped observations. Econometrica: J. Econ. Soc. 137–148 (1976)

Karian, Z.A., Dudewicz, E.J.: Fitting the generalized lambda distribution to data: a method based on percentiles. Commun. Stat. Simul. Comput. **28**(3), 793–819 (1999)

Kleiber, C.: A guide to the Dagum distributions. In: Chotikapanich, D. (ed.) Modeling Income Distributions and Lorenz Curves. Economic Studies in Equality, Social Exclusion and Well-Being, vol. 5, pp. 97–117. Springer, New York (2008). https://doi.org/10.1007/978-0-387-72796-7_6

Lyon, M., Cheung, L.C., Gastwirth, J.L.: The advantages of using group means in estimating the Lorenz curve and Gini index from grouped data. Am. Stat. **70**(1), 25–32 (2016)

McDonald, J.B.: Some generalized functions for the size distribution of income. Econometrica, 647–663 (1984)

Podder, N.: Distribution of household income in Australia. Econ. Rec. **48**(2), 181–200 (1972)

Prendergast, L.A., Staudte, R.G.: Exploiting the quantile optimality ratio in finding confidence intervals for quantiles. STAT **5**(1), 70–81 (2016a)

Prendergast, L.A., Staudte, R.G.: Quantile versions of the Lorenz curve. Electron. J. Stat. **10**(2), 1896–1926 (2016b)

Prendergast, L.A., Staudte, R.G.: A simple and effective inequality measure. Am. Stat. **72**(4), 328–343 (2018)

Prendergast, L.A., Staudte, R.G.: Decomposing the quantile ratio index with applications to australian income and wealth data. Eur. J. Pure Appl. Math. (2019, to appear 8-June-2019)

R Core Team: R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2017)

Shorrocks, A.F.: The class of additively decomposable inequality measures. Econometrica, 613–625 (1980)
Su, S.: Confidence intervals for quantiles using generalized lambda distributions. Comput. Stat. Data Anal. **53**(9), 3324–3333 (2009)
Tarsitano, A.: Estimation of the generalized lambda distribution parameters for grouped data. Commun. Stat. Theory Methods **34**(8), 1689–1709 (2005)
Theil, H.: Economics and information theory. Technical report (1967)
Wang, B.: bda: Density estimation for grouped data. R package version 5.1.6 (2015)
Zeileis, A.:. ineq: measuring inequality, concentration, and poverty. R package version 0.2-13 (2014)